Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

iBIG: An Integrative Network Tool for Supporting Human Disease Mechanism Studies

Jiya Sun [1,2], Yuyun Pan [1,2], Xuemei Feng [1], Huijuan Zhang [1,2], Yong Duan [3,*], Hongxing Lei [1,3,*]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3] UC Davis Genome Center and Department of Biomedical Engineering, Davis, CA 95616, USA

Received 15 June 2012; revised 10 July 2012; accepted 31 August 2012
Available online 4 January 2013

KEYWORDS

Network;
Functional module;
Disease mechanism;
Microarray;
Web server

Abstract Understanding the mechanism of complex human diseases is a major scientific challenge. Towards this end, we developed a web-based network tool named iBIG (stands for integrative BIolоGy), which incorporates a variety of information on gene interaction and regulation. The generated network can be annotated with various types of information and visualized directly online. In addition to the gene networks based on physical and pathway interactions, networks at a functional level can also be constructed. Furthermore, a supplementary R package is provided to process microarray data and generate a list of important genes to be used as input for iBIG. To demonstrate its usefulness, we collected 54 microarrays on common human diseases including cancer, neurological disorders, infectious diseases and other common diseases. We processed the microarray data with our R package and constructed a network of functional modules perturbed in common human diseases. Networks at the functional level in combination with gene networks may provide new insight into the mechanism of human diseases. iBIG is freely available at http://lei.big.ac.cn/ibig.

Introduction

Among many great challenges in the field of biological sciences, disease mechanism is of imminent relevance to every single person in the whole world. From the perspective of cellular networks, complex diseases are progressive transformations of the cellular network. For heritable diseases, the network is flawed at the very beginning. For chronic complex diseases, lifelong gene and environment interaction results in dynamic adjustment and, at certain points, breakdown of the network. Understanding the specific destruction of the network in specific diseases and at specific stages is the key starting point for subsequent design of rescue or remedial strategies.

Network analysis has been increasingly utilized in interpreting high throughput data. Networks can be constructed purely based on gene expression information, including transcriptional regulatory networks [1] and co-expression networks [2]. Networks can also be built upon prior knowledge of protein–protein interactions [3]. Several network tools have been

[*] Corresponding authors.
E-mail: duan@ucdavis.edu (Duan Y), leihx@big.ac.cn (Lei H).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

1672-0229/$ - see front matter
http://dx.doi.org/10.1016/j.gpb.2012.08.007

implemented as Cytoscape plugins, including BisoGenet [3], MIMI [4] and APID2NET [5], which mainly focus on protein interactomes. Network building is also provided by web services such as STRING (http://string-db.org/). Since cellular networks consist of various types of interaction and regulation, networks reflecting this complex scenario will provide better insight into the problem in hand.

In this work, we developed a network tool iBIG (stands for integrative BIoloGy), which incorporates information on both interaction and regulation. The main architecture consists of a client interface by HTML and JavaScript, a server-side script written in CakePHP and a MySQL database. The network visualization is implemented based on the Cytoscape Web application programming interface (API). An important R package ArrayPro (http://lei.big.ac.cn/download/open_download_page) is also provided for processing of microarray data and construction of networks based on functional gene sets. To illustrate this unique feature, an example of network perturbation in common human diseases is provided.

Design and implementation

iBIG architecture

iBIG is a client-server based application following the CakePHP framework, a popular MVC model. In MVC models, model (M) is used to access database and pass the result to control (C), which responds to client request and view (V) is set according to the result from control. The three main units in MVC models include client, server and database (**Figure 1**). When constructing networks of functional gene sets, ArrayPro can remotely access our database. Visualization can be used for networks generated by iBIG or other network tools. iBIG has been extensively tested on IE8 and Firefox.

Database design

A unique internal gene ID is used to represent every gene and its product. The IDs from public databases are converted to the internal gene IDs. Our integrated database mainly consists of two parts: gene interaction and gene annotation. The interaction data are further classified into primary interaction, secondary interaction and network regulation. Primary interactions include pathway interaction, protein complex interaction and general protein–protein interactions. Secondary interactions include gene-gene interaction, chromosome position interaction, transcription factor-target gene interaction and kinase-target interaction. To facilitate the understanding of regulatory relationships, the latter two together with microRNA-target gene interactions form the network regulation category.

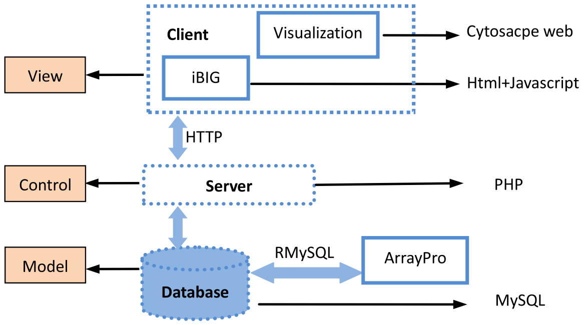


Figure 1 iBIG architecture

The client, server and database correspond to view, control and model in the MVC model, respectively. ArrayPro can access the database by RMySQL. Visualization is developed based on Cytoscape Web API.

Sources of the integrated database

Pathway interactions were collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) by R package 'KEGGSOAP' (available for downloading pathway gene information), WikiPathway (http://www.wikipathways.org/), NCI-Nature (http://pid.nci.nih.gov/), PathwayCommons [6] (this composite pathway database can be selected independently or further merged with other databases), Reactome (http://www.reactome.org/) and EHMN [7]. Protein complex interactions were collected from MIPS (http://mips.helmholtz-muenchen.de/genre/proj/corum). Protein–protein interactions were collected from HPRD [8], BioGrid (http://thebiogrid.org/), DIP [9], MINT [10], IntAct [11] and BIND [12]. Gene-gene interactions were downloaded from BioGrid. Chromosome position interactions were collected from the Molecular Signatures Database (MSigDB) of GSEA (http://www.broadinstitute.org/gsea/msigdb/index.jsp). Transcription factor-target gene interactions were collected from a recent paper [13]. Kinase-target interactions were collected from PhosphoSitePlus (http://www.phosphosite.org/). MicroRNA-target gene interactions were collected from TarBase [14], miRecords [15] and MicroCosm [15].

The network annotation includes the following information: pathway, protein complex, chromosome position, transcription factor, microRNA, kinase, epigenetics-related gene, housekeeping gene, tissue-specific gene, gene ontology (GO) biological process, GO molecular function and GO cellular component. We used the same data as in the interactions for pathway, protein complex, chromosome position, transcription factor, microRNA and kinase. Epigenetics-related genes were collected from GO and NCBI Entrez Gene. Housekeeping genes were collected from three papers [16–18]. Tissue-specific genes were collected from a recent paper [17]. Data for GO biological process, molecular function and cellular component were collected from MSigDB of GSEA.

Construction of gene networks

Construction of a gene network consists of several steps: (1) submit a gene list with one gene per line (gene symbol and Entrez gene ID are supported); (2) select databases for primary or secondary interaction (selection of primary interactions is mandatory while second interactions are optional); (3) set the network filtering strategy to reduce network complexity; (4) select databases for network regulation (optional) if the user wants to know about the upstream regulators and downstream targets; and (5) choose preferred network annotation to illustrate the functions of genes in the network. The generated network can be visualized online directly or downloaded in

XGMML format. The online visualization facilitates the interactive refinement of the network by modifying the selections.

Construction of networks with functional gene sets

ArrayPro is a supplementary R package mainly for microarray data processing, including data preprocessing, identification of differentially expressed genes (DEGs), functional enrichment analysis and construction of networks with functional gene sets. Networks with functional gene sets can be built by calculating the correlation among selected functional gene sets. In our recent work, ArrayPro has been applied to the investigation of network perturbation in Alzheimer's disease [19].

Calculation of relationship among functional gene sets

ArrayPro is an independent R package which can be downloaded from http://lei.big.ac.cn/download/open_download_page. One of the functions of ArrayPro is to build networks with functional gene sets instead of individual genes. The detailed procedures are described as follows. (1) Genes belonging to the relevant gene sets are selected. (2) A gene interaction network is constructed based on selected types of interactions. (3) For any pair of gene sets, such as gene sets A and B, the significance of node overlap $P_{\text{node\_overlap}}$ is calculated. (4) The significance of direct interaction between the two gene sets P_{direct} is also calculated (for two gene sets, genes from one gene set may interact with genes from the other gene set. This type of interaction is called direct interaction). (5) The combined probability of $P_{\text{node\_overlap}}$ and P_{direct} is calculated using Fisher's method [20]. In formula (1), $P_{\text{node\_overlap}}$ is the P-value of node overlap between gene sets A and B, P_{direct} is the P-value of direct interaction between gene sets A and B, and S is the score transformed from the combined probability. The P-value of S, which follows chi-square distribution with $2k$ degrees of freedom (k is the total number of variables to be combined, 2 in this case), is calculated by formula (2). If the P-value of S (P_S) between gene sets A and B is less than a given threshold such as 0.05, we consider the two gene sets functionally related and the *Score*, calculated by formula (3), is taken as the final score for the relationship between the two gene sets.

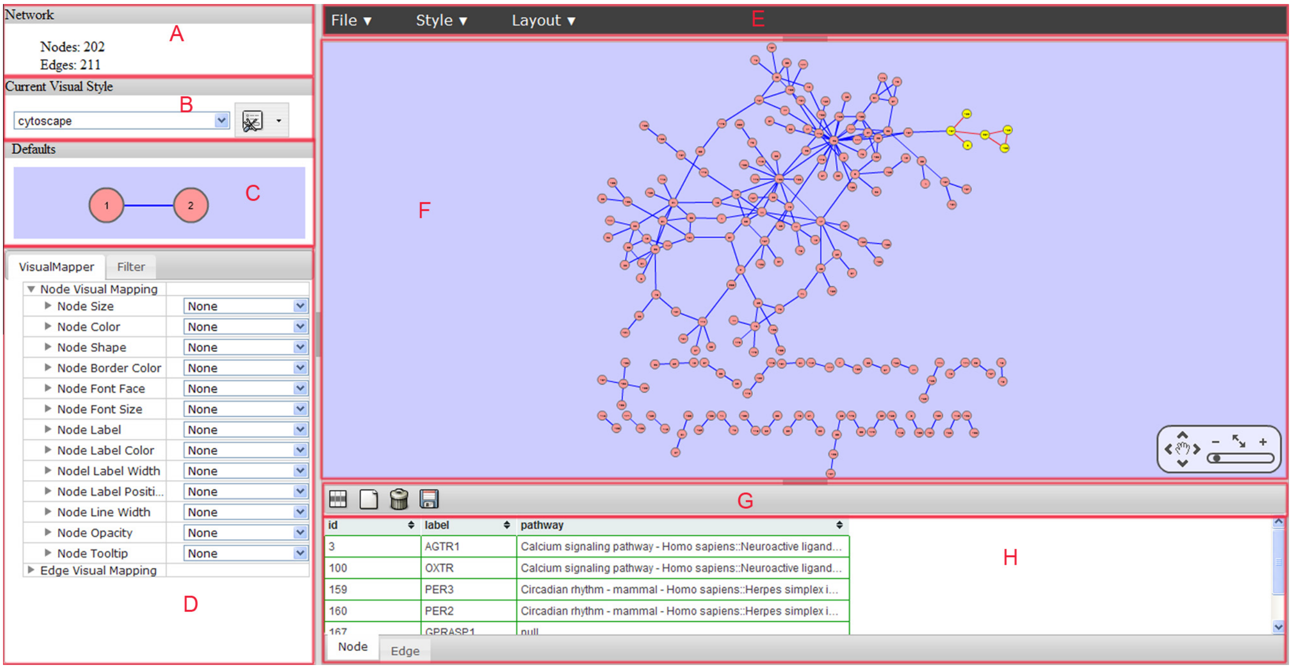
$$S = -2(\log(P_{\text{node\_overlap}}) + \log(P_{\text{direct}})) \tag{1}$$

$$S \sim \chi^2_{2k} \tag{2}$$

$$\text{Score} = -\log(P_s) \tag{3}$$

Network visualization

Standalone network tools such as Cytoscape [21] and VisANT [22] have been developed for the visualization of biological networks. Recently, a web-based visualization tool Cytoscape Web [23] has been developed, which uses flash technologies and provides a javascript API for developers. Implementation of our visualization tool is based on Cytoscape Web 0.7.4 release with the goal of mimicking the standalone Cytoscape. This convenient visualization tool (**Figure 2**) can be independently accessed at http://lei.big.ac.cn/visualization/start_



Network
Nodes: 202
Edges: 211

Current Visual Style
cytoscape

Defaults

Node Visual Mapping	
Node Size	None
Node Color	None
Node Shape	None
Node Border Color	None
Node Font Face	None
Node Font Size	None
Node Label	None
Node Label Color	None
Nodel Label Width	None
Node Label Positi...	None
Node Line Width	None
Node Opacity	None
Node Tooltip	None
Edge Visual Mapping	

id	label	pathway
3	AGTR1	Calcium signaling pathway - Homo sapiens::Neuroactive ligand...
100	OXTR	Calcium signaling pathway - Homo sapiens::Neuroactive ligand...
159	PER3	Circadian rhythm - mammal - Homo sapiens::Herpes simplex i...
160	PER2	Circadian rhythm - mammal - Homo sapiens::Herpes simplex i...
167	GPRASP1	null

Figure 2 Visualization interface

A. The numbers of nodes and edges of the network. **B.** The visual style container where users can create, rename and delete selected visual style. **C.** The window used to show the global visual style. **D.** The visual mapper panel and filter panel, where users can set visual style or filter nodes and edges according to specified attributes. **E.** The menus to import network, export network, import attribute and lay out network. **F.** The main window to display the network. **G.** The four buttons from left to right, including "show selected attribute", "create new attribute", "delete attribute" and "export selected attributes". **H.** The window used to show attributes of nodes and edges.

visualization, with which construction and operation of networks based on web browser can be easily achieved with the tactics from standalone Cytoscape.

Case study

Construction of a network for common human diseases

Microarray datasets were downloaded from NCBI gene expression omnibus (GEO) and EBI ArrayExpress. Fifty-four microarray datasets were used in this study, including 12 for cancer, 7 for neurological disorders, 29 for infectious and inflammatory diseases and 6 for metabolic diseases. Microarray data preprocessing, differential expression identification, enrichment analysis and construction of functional networks were all performed with ArrayPro. The microarray raw data (CEL files) was preprocessed with the GCRMA algorithm to get the expression values for every probe. Any probe sets with a call value of less than 10% returned by mas5calls function in affy package were removed. Then, probe sets were mapped to Entrez Gene ID. Any probe sets not mapped to known genes were also removed from further analysis. If there are multiple probe sets mapped to the same gene, we averaged their expression values as the expression of the gene. Differential expressional genes were identified by the FC-based RankProd algorithm. Enrichment analysis was based on gene sets including EHMN, KEGG, NCI and GO from GSEA. For every disease group, 60 functional terms (gene sets) were selected according to the enrichment score. A total of 240 functional terms from the four disease groups were merged together, which resulted in 117 nodes (functional terms) for the functional network. The functional network was constructed by ArrayPro based on the HPRD database. Interactions with $P < 0.01$ were considered significant.

Network perturbation in common human diseases

One of the unique features of iBIG is the construction of networks with functional modules. This feature can facilitate the understanding of the investigated biological problem at a higher level compared to gene networks. In our recent work, we have used this functionality in the investigation of pathogenesis of Alzheimer's disease [19]. Here we demonstrate this functionality by constructing a functional network perturbed in common human diseases. The most significantly perturbed functional modules in each of the four classes of diseases were selected and merged together. The connectivity among this set of 117 functional modules was calculated again by ArrayPro and the network was thus constructed (Figure 3). Here we briefly describe the relevance of this network to the mechanism of human diseases.

Many of the uniquely-perturbed functions in a specific disease class are consistent with the current knowledge. For example, cell cycle, DNA replication and p53 pathway (KEGG) are perturbed only in cancer, while transmission of nerve impulse, synaptic transmission, nervous system development, long term potentiation, long term depression, axon guidance and gap junction are perturbed only in neurologic diseases. Therefore, other uniquely-perturbed functions may also play important roles in the specific class of diseases. For

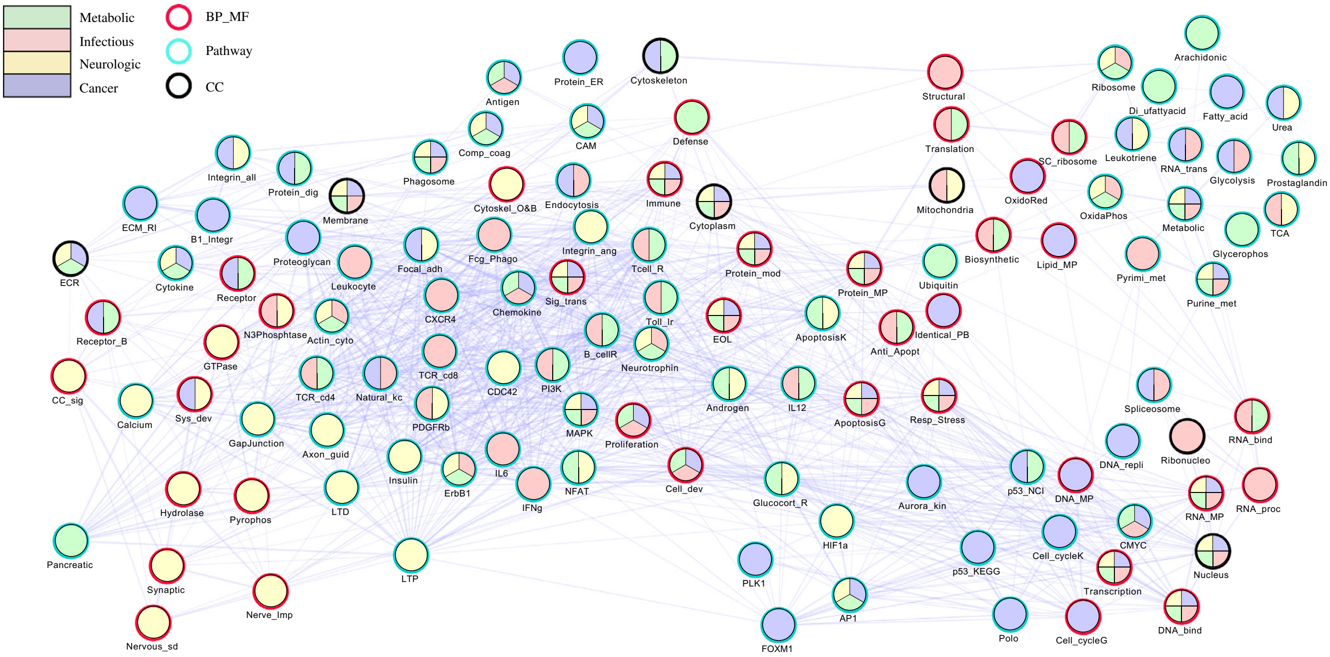


Figure 3 Network perturbation in common human diseases
The most significantly perturbed functional modules in four disease classes are merged together to form a comprehensive network perturbed in common human diseases, including cancer, infectious diseases, neurologic diseases and metabolic diseases. "BP_MF" stands for biological process and molecular function in gene ontology (GO), while "CC" stands for cellular component in GO. "Pathway" stands for curated pathways in KEGG and NCI pathway databases.

cancer, lipid metabolism and protein processing in endoplasmic reticulum (ER) are two of the less-known factors. Lipid metabolism is involved in membrane formation and energy production which are both critical for cell proliferation [24]. Cell proliferation will also have a different demand on the protein folding and recycling in ER. For neurologic diseases, more attention may be paid to calcium signaling, insulin signaling and HIF1alpha transcriptional regulation. Calcium signaling is of great importance to the maintenance of normal neurologic activities. Insulin signaling is involved in nutrient sensing and adjustment of cellular activity [25,26]. HIF1alpha transcriptional regulation is involved in oxygen sensing and cell fate decisions. In our recent work, we have proposed that the cause of Alzheimer's disease is the prolonged low supply of oxygen and nutrients in the brain [19]. For infectious diseases, the unique perturbation of ribonucleoprotein complexes is of particular interest. The significant perturbation of this functional module likely reflects the enhanced translational activity in the ribosome.

On the other hand, some functional modules are perturbed in three classes of diseases but not in the fourth class. For example, the extracellular region and cell adhesion molecules are not significantly perturbed in infectious diseases, consistent with the transcriptome measurement on blood for this class of disease which lacks tight cellular connection as in other tissues [27]. The non-significant perturbation of the complement and coagulation cascade is a little surprising. The complement and coagulation cascade is involved in immune response and blood clotting [28]. Up-regulation of this functional module is only observed in Tuberculosis (data not shown), leading to overall non-significant perturbation in this disease class. The non-significant perturbation of the AP1 transcriptional network in infectious disease may also deserve further investigation. AP1 functions in many cellular activities including cell cycle proliferation and apoptosis [29]. The dysregulation of the AP1 transcriptional network has been reported in cancer and neurological diseases, while its connection with infectious diseases has rarely been reported. For cancer, the ribosome is not significantly perturbed. This may indicate non-significant overall perturbation of translational activity in cancer despite the significant dysregulation of cell cycle. For neurologic diseases, CMYC pathway and antigen processing and presenting are not significantly perturbed. CMYC is involved in apoptosis under certain conditions, but this may not be relevant to neurologic disorders. Antigen processing and presenting is involved in the immune response process. The lack of significant perturbation of this functional module may indicate a non-significant immune response in neurologic diseases. An interesting observation is the lack of functional modules significantly perturbed in three disease classes but not in metabolic diseases, likely due to the less consistent and specific perturbation among metabolic diseases.

In addition, hidden links between a pair of disease classes can be revealed on this network. Cancer and neurological diseases shared significant perturbation of the integrin family cell surface interaction and focal adhesion. This integrin-focal adhesion axis is involved in cell proliferation and apoptosis, which are prominent features of the two disease classes [30]. Cancer and infectious diseases shared significant perturbation of endocytosis. Endocytosis is an important defense mechanism against pathogens. It has also been found that endocytosis is involved in other functions including a variety of signaling events [31]. Neurological diseases and infectious diseases shared significant perturbation on energy metabolism related functional modules including mitochondria and TCA cycle. This may reflect the special energy requirement in these two disease classes. In addition, cancer and metabolic diseases shared significant perturbation of receptor activity and receptor binding. Neurologic diseases and metabolic diseases shared significant perturbation of hormone mediated signaling pathways including glucocorticoid receptor signaling pathway and androgen mediated signaling. Infectious diseases and metabolic diseases shared significant perturbation of translation related functional modules. Due to the heterogeneous nature of metabolic diseases, it is not immediately clear how those intersections are related to the disease mechanism.

Network analysis has been widely applied to the investigation of human disease mechanisms. In most of the studies, the major focus is on the gene network. Here we provide the functional network as a complementary view of the studied biological problem. Gene networks can provide detailed information on gene–gene interaction and regulation, while functional networks can provide a global view of the cellular transformation. The combination of these two types of networks will provide more comprehensive understanding of the studied problem including the disease mechanism. Currently we are applying this strategy to the in-depth investigation of cancer and Alzheimer's disease.

Conclusion

In summary, iBIG is a new tool for network construction and visualization. Distinct features include classification of interactions, web-based visualization and networks of functional gene sets. The web-based visualization provides a convenient way to refine networks interactively. Future development of iBIG will include integrating more functional data and further improvement the network visualization. Because our remote database is based on several external databases, we plan to update it manually and periodically (twice a year).

Authors' contributions

JS conceived the design, wrote major part of the iBIG code and wrote the first draft. YP conducted the analysis on disease network. XF wrote the ArrayPro code. HZ wrote part of the iBIG code. YD conceived the design and revised the manuscript. HL conceived the design and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

None declared.

Acknowledgements

This work is supported by research grants from National Institute of Health to YD (Grant No. GM79383 and GM67168), Natural Science Foundation of China to HL (Grant No. 30870474) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry to HL.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.08.007.

References

[1] Chen X, Chen M, Ning K. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics 2006;22:2952–4.

[2] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.

[3] Martin A, Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de-Cossio J, Bringas R. BisoGenet: a new tool for gene network building, visualization and analysis. BMC Bioinformatics 2010;11:91.

[4] Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR, Jagadish HV, et al. Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics 2009;25:137–8.

[5] Hernandez-Toro J, Prieto C, De las Rivas J. APID2NET: unified interactome graphic analyzer. Bioinformatics 2007;23:2495–7.

[6] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway commons, a web resource for biological pathway data. Nucleic Acids Res 2011;39:D685–90.

[7] Hao T, Ma HW, Zhao XM, Goryanin I. Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics 2010;11:393.

[8] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. Nucleic Acids Res 2009;37:D767–72.

[9] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30:303–5.

[10] Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, et al. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 2010;38:D532–9.

[11] Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010;38:D525–31.

[12] Bader GD, Hogue CW. BIND – a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 2000;16:465–77.

[13] Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin JB. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res 2010;38:e120.

[14] Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA 2006;12:192–7.

[15] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. MiRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009;37:D105–10.

[16] Zhu J, He F, Song S, Wang J, Yu J. How many human genes can be defined as housekeeping with current expression data? BMC Genomics 2008;9:172.

[17] She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen R. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics 2009;10:269.

[18] Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 2006;7:31.

[19] Sun J, Feng X, Liang D, Duan Y, Lei H. Down-regulation of energy metabolism in Alzheimer's disease is a protective response of neurons to the microenvironment. J Alzheimers Dis 2012;28:389–402.

[20] Fisher RA. Questions and answers #14. Am Stat 1948;2:30–1.

[21] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.

[22] Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C. VisANT: data-integrating visual framework for biological networks and modules. Nucleic Acids Res 2005;33:W352–7.

[23] Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape web: an interactive web-based network browser. Bioinformatics 2010;26:2347–8.

[24] Damrauer SM, Studer P, da Silva CG, Longo CR, Ramsey HE, Csizmadia E, et al. A20 modulates lipid metabolism and energy production to promote liver regeneration. PLoS One 2011;6:e17715.

[25] Rosenberg SS, Spitzer NC. Calcium signaling in neuronal development. Cold Spring Harb Perspect Biol 2011;3:a004259.

[26] Porte Jr D, Baskin DG, Schwartz MW. Insulin signaling in the central nervous system: a critical role in metabolic homeostasis and disease from *C. elegans* to humans. Diabetes 2005;54:1264–76.

[27] Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 2006;147:126–32.

[28] Amara U, Rittirsch D, Flierl M, Bruckner U, Klos A, Gebhard F, et al. Interaction between the coagulation and complement system. Adv Exp Med Biol 2008;632:71–9.

[29] Shaulian E, Karin M. AP-1 as a regulator of cell life and death. Nat Cell Biol 2002;4:E131–6.

[30] Shibue T, Weinberg RA. Integrin beta1-focal adhesion kinase signaling directs the proliferation of metastatic cancer cells disseminated in the lungs. Proc Natl Acad Sci U S A 2009;106:10290–5.

[31] von Zastrow M, Sorkin A. Signaling on the endocytic pathway. Curr Opin Cell Biol 2007;19:436–45.